基于论文摘要和引文文本语料的突破性研究特征词识别*

■ 杨雪梅[1] 王雪[1] 杜建[2] 唐小利[1]

[1] 中国医学科学院医学信息研究所 北京 100005 [2]北京大学健康医疗大数据国家研究院 北京 100191

摘 要：[目的/意义]基于作者对自身研究的描述性评价和后续研究者的评论性引用视角，利用摘要和引文语料提取突破性研究的特征词，从而了解突破性研究的摘要和引文语料特征以帮助对于突破性研究的识别。[方法/过程]选取 *Science* 评选为“Breakthrough of the Year”的关键文献和 Nobel Prize 获得者的“key publications”作为突破性研究语料数据，整合论文的摘要和引文语料进行特征词提取。特征词提取中，首先利用 Stanford CoreNlp 工具对语料进行分词及词频统计，并结合专家意见提取特征词元。然后将特征词作为种子词，利用医学文本的语义关系对特征词进行语义拓展。最后通过查全率和查准率进一步对比摘要和引文的特征词拓展前后的检索识别效果。[结果/结论]突破性研究语料中遴选出8个摘要语料的特征词元和8个引文语料的特征词元。特征词检索识别中，摘要和引文的拓展特征词的查全率最高，引文特征词的查准率最高，引文拓展特征词的查全率和查准率综合效果较好。

关键词：突破性研究 特征词 摘要文本 引用语句

分类号：G250

DOI：10.13266/j.issn.0252-3116.2020.11.014

“创新驱动发展”已成为我国加快推动经济发展方式转变的战略举措，其中，能够带来产业技术架构与组件双重变革和市场颠覆的突破性技术创新是创新驱动发展的破局之举。突破性研究的早期发现可以直接推动突破性技术创新，如果能够在研究早期识别出突破性研究，就能够促进突破性创新研究的部署，加快推动突破性技术创新的进程。突破性研究的早期识别对于我国科技创新强国的建设具有重要意义。

本文从特征词发现的角度，挖掘突破性研究的摘要和引文语料的语义特征，利用提取的特征词帮助突破性研究文献能在大规模文献数据集中被识别和发现，并为突破性研究的精准检索限定候选突破性文献集，实现突破性研究的早期识别，为突破性研究的部署战略提供理论支撑。

1 相关研究

1.1 突破性研究识别

目前对突破性研究进行识别的相关研究主要集中在科学计量学领域，研究方法可以分为两类：一类是基于文献计量学的特异性指标识别突破性研究。E. Garfield 通过观察科学论文随时间的影响，利用简单的引文计数初步识别出部分科学发现[1]。I. V. Ponomarev 等[2]基于出版物引用动态，结合定量方法，早期发现、识别候选突破性论文。Y. H. Huang 等[3]基于文章的引用路径(引文链)，发现突破性研究的出现会引发引文链的破裂，因此提出用“破裂分数”识别生物医学等领域的突破性研究。除了引文的相关指标，一些研究还综合考虑了作者合作、时间延迟承认指数等指标。科睿唯安与美国国家癌症研究所的一项联合研究[4]中，将文献的出版时间、共同作者网络及其他字段特征纳入随机森林模型，结合主题专家的遴选，实现在论文发表后能尽早识别候选的突破性论文。杜建等[5]提出使用延迟承认指数和被专利引用两个指标识别变革性研究。

还有一类方法是利用评价性语句的语言特征识别突破性研究。众多学者因引文语料中包含大量有价值的信息，采用引用内容分析或施引语句分析方法来识别

* 本文系国家社会科学基金项目“基于科学与技术交叉模型的创新前沿识别方法与应用研究”(项目编号：18BTQ064)和中国医学科学院医学与健康科技创新工程“医学科技创新评价与卫生服务体系构建研究”(项目编号：2016-I2M-3-018)研究成果之一。

作者简介：杨雪梅(ORCID：0000-0002-2927-4166)，助理馆员，硕士；王雪(ORCID：0000-0001-6852-1791)，硕士研究生；杜建(ORCID：0000-0002-7621-9995)，副研究员，博士；唐小利(ORCID：0000-0001-6946-3482)，研究馆员，硕士，通讯作者，E-mail：tang.xiaoli@imicams.ac.cn。

收稿日期：2019-10-23 修回日期：2019-12-28 本文起止页码：125-132 本文责任编辑：杜杏叶

突破性研究。D. R. Radev 等从引文语义挖掘的角度，利用施引语句的描述总结被引用论文的“贡献点”[6]。H. Small 选择出现线索词“discover * ”的施引语句及其对应的参考文献，利用施引语句的语义特征进行机器学习试验，自动判断参考文献是否为“科学发现”[7]。

以上的研究均以引用数据视角开展重大发现的识别，由于作者在引用参考文献时动机复杂，基于引用次数和引用路径的指标研究，存在简单的引用计数和引用关系并不能提供足够的信息来进行准确的识别的弊端。因此基于引用语句特征的识别虽然重点突出了后续研究者对相关研究的评价，但每条引文并不能全面概括作者的研究成果，因此，基于语句特征的识别还应结合论文作者对自身研究的评价。此外，H. Small 利用线索词限定候选突破性研究文献为突破性研究识别提供了新的思路，但是单一的线索词难免会将一些潜在的突破性研究文献排除在外。

1.2 文本特征提取

文本特征提取是自然语言处理的基本步骤之一，已经在文本分类、文本标引及文本检索等领域得到广泛应用。文本特征提取的主要思想是首先构建一个评估函数，然后通过这个函数计算特征词条的权重，接着对特征词的权重进行排序，选取前 n 个特征为最终的特征词集合中的子集[8]。

目前文本特征提取方法主要分为两类：基于统计的方法和基于语义的方法[9]。基于统计方法的评估函数是类间不相关的，即词与词之间默认是没有联系的。在词频统计的方法中，TF-IDF (term frequency-inverse document frequency) 方法一直是研究的重点，并能表现较好的提取效果。TF-IDF 由 G. Salton 等于 1988 年首次提出[10]，其中 TF 称为词频，用于计算该词描述文档内容的能力；IDF 称为反文档频率，用于计算该词区分文档的能力。谷俊等[11]利用 ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) 分词系统抽取专利文档的词元，将获取的词元通过改进的 TF-IDF 模型进一步筛选热点词元，最后由专家人工判定出有效的新技术术语。基于统计的方法具有过程简单、结果直观的特点，但忽略了文本中词义关系和语义特征，导致特征词提取不够全面。

基于语义方法的评估函数是类间相关的，即建立在语义理解的基础上，通过对上下文关系的提取构建语义网络。目前提取效果较好的是 T. Mikolov 等提出的 Word2vec 模型[12]，该模型可以根据上下文关系以词向量的形式提取表达语义，并在向量空间内将词的向量按相似性进行分组。C. Chen 在研究科学出版物中代表不确定性的特征词时，利用 Word2vec 将文本进行向量化处理，进而根据一系列的种子词选择不确定性的特征词[13]。总的来说，基于语义特征提取的方法，能够有效地提高文本信息特征提取的准确性，但是高准确率仅体现在已定义的类别间，而对于尚未定义的域外类别，类间相关评估函数的选择效果也不理想。

2 特征词提取方法

基于相关学者对突破性研究识别及特征词提取的研究，本文在后续研究者对文献评价的基础之上，整合作者对自身研究的评价，利用词频统计和语义拓展的方式对已知突破性研究的摘要和引文语料进行特征词提取。从“自评 + 他评”的视角，更全面、更准确地提取突破性研究的特征词。

基于论文摘要和引文文本语料的突破性研究特征词识别主要分为四步：数据来源与预处理、基于词频统计的特征词元提取、基于语义的特征词元语义拓展、特征词提取效果评估，提取方法的框架见图 1。

2.1 数据来源与预处理

在已知突破性研究文献的选择过程中，需要明确突破性研究的定义，限定已知突破性研究的文献选取范围，方便数据获取。目前对突破性研究 (Breakthrough Discoveries) 尚无公认的定义，I. V. Ponomarev 等在识别已知突破性论文和高被引论文的典型引用模式过程中，发现突破性论文获得较多的引用而且能够为当下的研究提供新的方向[14]。在 *Science* 评选的年度科学突破 (Breakthrough of the Year) 中，其新闻副主编 R. Coontz 表示：科学突破应该是起到两种作用中的一种，或者解决了一个人们长时间苦思冥想的问题，或者为许多新研究开启了大门[15]。本文将突破性研究界定为在渐进式的研究中做出的重大发现，或在原有研究基础上的颠覆、变革，并为研究提供了新方向。

Science 评选的“年度科学突破” (Breakthrough of the Year) 被广泛认为是科学领域的最高成就之一[16]，这些被评选的科学突破符合本文界定的突破性研究定义。此外，同样作为科学领域最高成就之一的 Nobel Prize，其获奖者均是在物理、化学、医药等领域具有重大发现的学者。因此本文选择在生物医学领域，被 *Science* 评选为年度科学突破 (Breakthrough of the Year) 的参考文献及 Nobel Prize 获得者的“代表作” (Key Publications) 作为突破性研究特征词提取语料。文献数据在 *Science*、Nobel 官网检索获取。

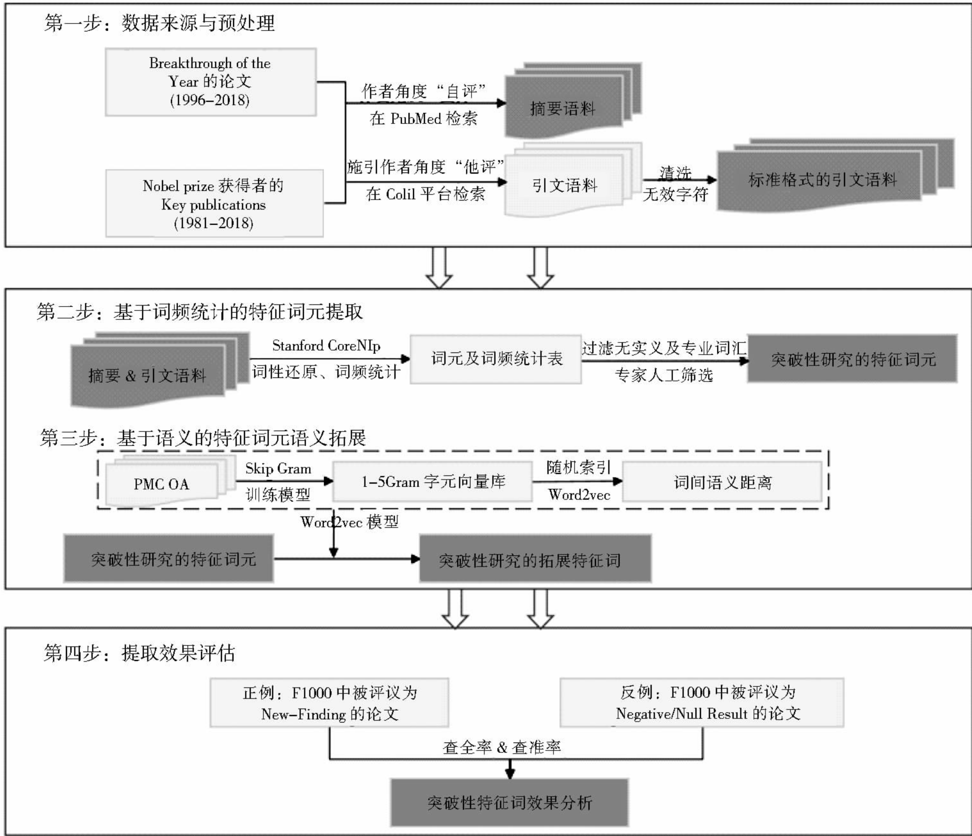


图 1　突破性研究特征词提取方法框架

基于作者对自身研究的描述性评价和后续研究者的评论性引用视角，分别选择论文摘要和引文文本开展突破性研究的特征词识别。论文摘要语料主要利用论文的 PMID 在 PubMed 数据库获取；引文文本语料通过文章的 PMID 在 Colil 平台检索[17]，该平台是日本国家生命科学数据库中心基于 PMC-OAS 开发，输入文章 PMID 能够直接批量获取文章的引文文本信息[18]。通过 PMID 获取的论文摘要和引文文本语料均为结构化数据，但是由于原始语料是通过网络、人工录入、软件识别转换等方式加工存储，文本常存在符号、格式不规范等问题，直接使用会影响数据统计、赋码等多个环节。因此在特征词提取前，语料数据需要剔除无效字符、参考文献、网址等非标准化字符。

2.2　基于词频统计的特征词元提取

突破性研究语料特征词选取的重点是选出多篇文献共同提到的特征词，无需考虑反文档频率，因此 TF-IDF 的方法并不适用于此处的特征词元选取。在筛选特征词元过程中，本文选择传统的词频统计方式，使用 Stanford CoreNlp 工具对语料进行分词及词频统计[19]，提高词频统计的准确性。Stanford CoreNlp 工具获取语料词频的步骤为：分词－词形还原－基于句法的词性标注－词频统计，在此基础上过滤标点及属性为 CD（纯数，基数）的词，减少标点及数字带来的噪音。图 2 是以“The sulfur atom is supplied by a separate cluster in the enzyme.”为语料示例展示的词频统计过程：

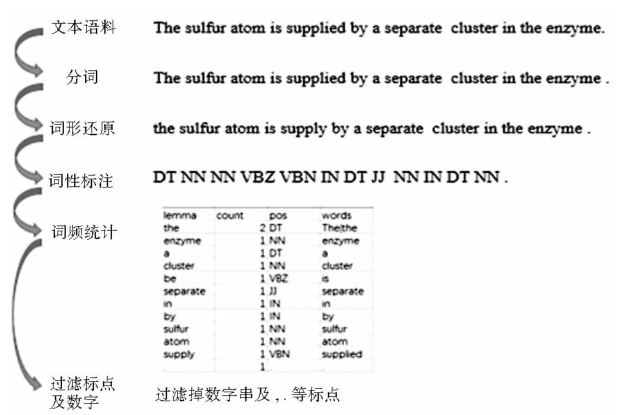


图 2　语料词频统计示例

2.3　基于语义的特征词元语义拓展

在根据 NLP 方法获得了特征词后，综合考虑 PMC 数据库中所有医学文献的语义关系进行特征词的语义拓展。Word2vec 主要采用 CBOW（Continuous Bag-of-Words Model）和 Skip-Gram（Continuous Skip-Gram Mod-

el)模型，其中 Skip-Gram 适用于更大规模的语料集[20]，因此本文选择 kip-Gram 模型。词义拓展中，首先基于 PMC Open Access（PMC OA）文章的所有文本内容，Skip-Gram 训练模型收集 1-grams 到 5-grams 的滑动窗口语料[21]。进而将获取的 N-gram 语料构建词元向量，得到包含构建分布相似性模型所需的所有显著信息的 N-gram 库。在此基础上利用随机索引方法[22]对语料库中所有上下文窗口中的单词的索引向量求和，以获得给定单词的向量空间。最后将编码过给定单词带入神经网络进行训练，通过计算给定单词间的余弦距离确定单词的矢量距离。单词矢量距离就是词义拓展的核心，矢量距离越近代表词义越相近。

通过以上方法完成 PMC OA Word2vec 模型的构建，模型构建的流程图见图 3。使用该模型进行词义拓展时，只需某个词输入到模型中，即可输出与这个词义更接近的词。

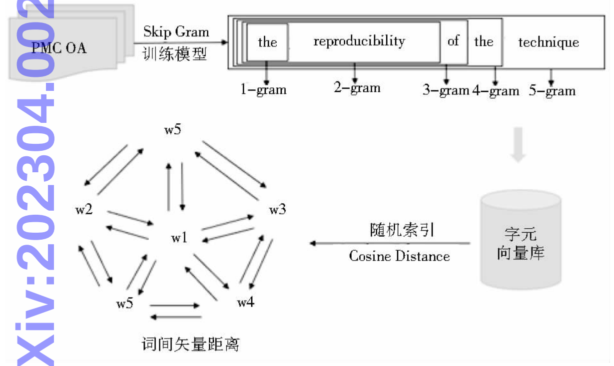


图 3　PMC OA Word2vec 模型构建的流程图

2.4　提取效果评价方法

在信息检索领域内，查全率和查准率[23]是反映检索效果的重要指标，因此可采用查全率与查准率两个指标判断特征词提取的效果。查准率是衡量检索信息噪声比的指标，即检出的相关文献量与检出的文献总量的百分比。查全率是衡量从文献集合中检出相关文献成功度的指标，即检出的相关文献量与相关文献总量的百分比。在突破性研究特征词检索效果评价中，查全率 = TP/（TP + FP），查准率 = TP/（TP + FN），字母表达的含义如表 1 所示：

表 1　查准率与查全率指标中字母含义说明

分类	突破性研究的文献	对照组文献
检出的文献	TP（真正例）	FN（假正例）
未检出的文献	FP（假反例）	TN（真反例）

正例与反例选择突破性研究文献和不具有突破创新的文献，数据分别来源于 Faculty of 1000（简称 F1000）数据库中被评议为 New-Finding 和 Negative/Null Result 的论文。其中被评议为 New-Finding 的文章中作者展示了新颖的数据、模型等，可认为是突破性的研究发现，被评议为 Negative/Null Result 的论文中作者得到阴性结果，或未展示有价值的结果，可以认为是不具有突破创新、价值较低的文献。

3　突破性研究的特征词提取

3.1　突破性研究数据获取及预处理

考虑数据的可获取性，本文纳入 1981 – 2018 年生物医学领域 Nobel prize 获得者的 Key Publications，及 1996 – 2018 年的 Breakthrough of the Year 论文。在 Science 及 Nobel 官网检索得到 Breakthrough of the Year 的论文 556 篇论文，Nobel prize 获得者的 Key Publications 相关文献 103 篇，二者去重得到 648 篇突破性研究文献。突破性研究摘要语料通过 PubMed 数据库进行检索得到 467 条语料。突破性研究的引文语料通过 Colil 平台检索得到 135 526 条语料，清洗后剩余 131 767 条引文语料。

3.2　特征词元筛选与提取

对突破性研究的摘要和引文语料进行词频统计，在摘要语料中提取 7 058 个词，累计词频为 54 394；引用语料中得到 70 995 个词，累计词频为 3 184 578。两组语料中，均以 NN（名词）、JJ（形容词）和 VB（动词）这三类词性为主，但是在占比上存在一定差异。一般而言，自然语言处理的结果随着语料的增加而逐渐稳定，因而引文语料的词性占比更具稳定性。具体的词性占比情况见图 4，内环为突破性研究摘要语料，外环为突破性研究的引文语料。

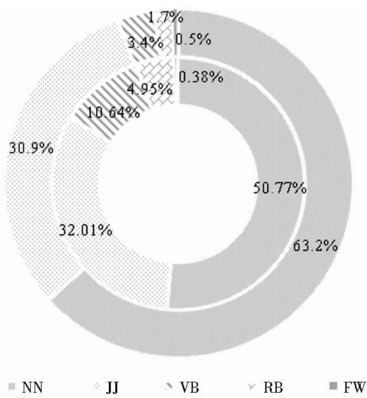


图 4　摘要和引文语料的词性占比

摘要语料和引文语料特征词选取中，选择词频排序 Top500 的词汇进行筛选，剔除 Mesh 词表中的医学

领域专业词汇,以及与突破性评价无关的词汇,得到部分词汇作为候选特征词。特征词元的进一步筛选,需要专家对摘要和引文文本进行研读,根据语料中相关学者的表述方式遴选。例如,在引文语料中经常出现下述表现形式,以 discover、since 等词汇突出表现被引论文作者的重大发现:

RNA interference (RNAi) was first discovered in C. elegans and has since been widely used to

Following the initial discovery of RNAi, a variety of small RNA-mediated silencing phenomena have been uncovered.

在遴选过程中,三位图书情报领域的专家以“背对背”的方式查看摘要和引文中包含候选特征词的语料。通过专家遴选,两名及以上的专家认为词元能表征突破性,则确认为特征词元。最终得到摘要语料的 8 个特征词元:new、novel、potential、key、change、evidence、basis、base,以及引文语料的 8 个特征词元:change、first、potential、new、novel、since、discovery、discover,表 2 展示了摘要和引文语料的特征词及其提取词元前对应的词:

表 2　摘要和引文语料的特征词汇

摘要语料的特征词元		引文语料的特征词元	
词元	提取词元前的词	词元	提取词元前的词
new	new	change	Changes | Changing | change | changed | changes | changing
novel	Novel | novel	first	First | first
potential	Potential | potential | potentials	potential	Potential | potential | potentials | potentialities | potentiality | potentialize
key	Key | key	new	New | new
change	Changes | change | changed | changes	novel	Novel | novel
evidence	Evidence | evidence | evidenced	since	Since | since
basis	bases | basis	discovery	Discoveries | discovery
base	Based | base | based	discover	Discovered | discover | discovered | discovering | discovers

3.3　词义拓展分析及可视化

通过将摘要和引文语料获取的特征词元输入到 Word2vec 模型(摘要词元的词义拓展参数设置为“new,novel,potential,key,change,evidence,basis,base” －n 50,引文词元的词义拓展参数设置为“change,first,potential,new,novel,since,discovery,discover” －n 50)表 3 显示了在 Word2vec 模型中与特征词 key、discovery 密切相关的前 10 个单词,这些词被认为是扩展的候选词。

表 3　特征词拓展候选词示例

特征词:key		特征词:discovery	
拓展词	矢量距离	拓展词	矢量距离
findslot	0.623 147	discoveries	0.485 044
keys	0.531 986	discovered	0.484 780
signature	0.478 537	discoverer	0.399 614
major	0.399 389	creation	0.387 769
pivotal	0.385 623	institute	0.382 202
unique	0.376 219	rediscovery	0.381 493
specific	0.359 634	findings	0.358 403
primary	0.349 635	latest	0.357 655
particular	0.347 595	foundation	0.343 410
signatures	0.338 006	contributions	0.325 022

通过专家评议并结合词语应用场景进一步筛选得到摘要拓展特征词 30 个,引文拓展特征词 36 个。在摘要语料中,摘要拓展特征词的共现关系见图 5,由于拓展特征词 changes、changing、credible 在摘要语料中不存在,所以共现图中有 27 个节点。图 5 中节点大小代表语料中出现频次,线的粗细代表节点的共现强度。从图中可知,based、changes、findings 等 19 个词的出现频率较高,且词汇之间的共现次数较多,尤其是 new 的出现频次最多,且与 evidence 的共现次数较多。进一步说明拓展特征词的有效性,且在摘要语料中多个特征词经常共同出现。

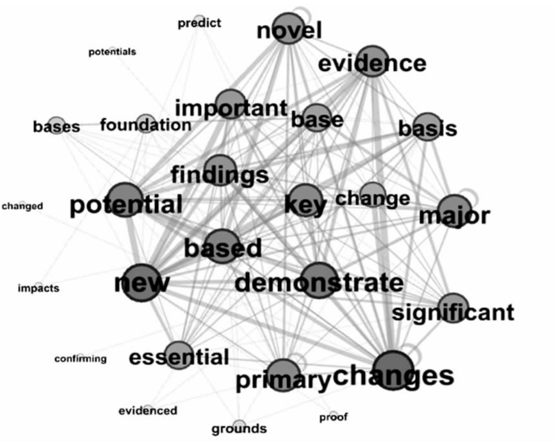


图 5　摘要语料中摘要拓展特征词的共现关系

在引文语料中，引文拓展特征词的共现关系见图 6，由于拓展特征词 devote 在引文语料中不存在，所以共现图中有 35 个节点。从图 6 中可知，大多数特征词的出现频率都较高，词共现网络更紧密，first 和 since 出现频率均较高，而且二者之间展示了极强的共现关系，在引用语料中多次出现 since…first…的句型。此外 since 与 discovery、first 与 discovered 的共现次数也较多，说明在引用语料中普遍多个特征词共同出现。



图 6　引文语料中引文拓展特征词的共现关系

3.4　突破性研究特征词的提取效果分析

特征词提取效果采用查全率与查准率进行评估，效果分析中选择 F1000 数据库中 5 次以上被评议为 New-Finding 的 183 条摘要语料和 1895 条引文语料作为正例，所有仅被评议为 Negative/Null Result 的 125 条摘要语料和 1840 条引文语料作为反例。通过特征词反向检索两组文献语料，计算摘要特征词 – 摘要拓展特征词、引文特征词 – 引文拓展特征词、摘要和引文特征词 – 摘要和引文拓展特征词在突破性研究检索过程中的查全率和查准率，具体情况见图 7。从图 7 中可知摘要和引文的拓展特征词的查全率最高，达到 94.54%，引文特征词查准率最高，达到 70.77%，查准率和查全率综合效果较好的是引文的拓展特征词。在突破性研究检索识别过程中可根据查全和查准的需求选择不同的特征词。

4　结论与展望

本文以生物医学领域为例，选取 *Science* 评选为“Breakthrough of the Year”的关键文献和 Nobel Prize 获

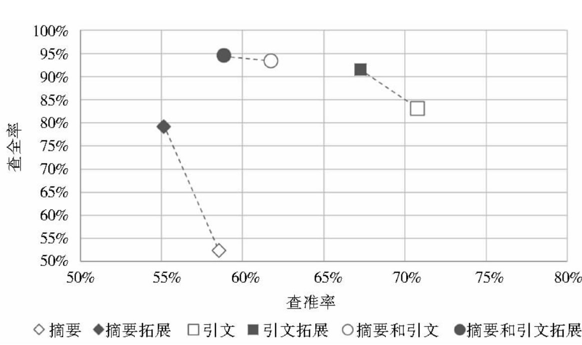


图 7　不同类别特征词的查全率与查准率

得者的“Key Publications”作为突破性研究语料数据，整合论文的摘要和引文语料，利用基于统计和基于语义的特征提取方法进行特征词提取。在研究中发现，不管是摘要语料还是引文语料，科技论文的语料词性整体上的构成相差无几。利用 Stanford CoreNlp 工具对摘要和引文语料进行分词及词频统计，得到 8 个摘要语料的特征词元：new、novel、potential、key、change、evidence、basis、base，8 个引文语料的特征词元：change、first、potential、new、novel、since、discovery、discover。利用 Word2vec 的方法进行特征词的语义拓展，最终得到 30 个摘要语料的拓展特征词，36 个引文语料的拓展特征词。通过共现分析，无论是在论文摘要还是引文文本中，特征词普遍共同出现，尤其是在引文语料中高频出现 since…first…的句型。

在特征词提取效果评价中，选择在 New-Finding 和 Negative/Null Result 的论文语料进行特征词的反向检索，检索结果显示摘要和引文的拓展特征词的查全率最高，但引文特征词的查准率最高，达到 70.77%，查准率和查全率综合效果较好的是引文的拓展特征词。相关学者在利用特征词检索突破性研究文献的过程中，可根据查全率和查准率的不同需求选择不同的特征词，特征词表见表 4。

本文利用突破性语料提取特征词，通过特征词识别突破性研究文献的查全率能够达到 90% 以上，但仅仅依靠特征词识别突破性研究的识别准确率还远远不够。识别出的特征词作为识别突破性研究的第一步，为突破性研究的识别初步划出研究的文献范围，后续的突破性研究识别中可以利用机器学习方法结合引文与摘要的整体语义信息，深入挖掘突破性研究语义特征，能够在候选突破性研究文献中准确识别突破性研究。

表 4 不同类型的特征词词表

特征词类型	检索范围	检索式	检索特点
引文特征词	引文	@ Citation(changes OR changing OR change OR changed OR first OR new OR potential OR potentials OR potentialities OR potentiality OR potentialize OR since OR novel OR discoveries OR discovery OR discovered OR discover OR discovering OR discovers)	查准率最高
引文拓展特征词	引文	@ Citation(affect OR beginning OR change OR changed OR changes OR changing OR contributions OR devote OR discover OR discovered OR discoverer OR discoveries OR discovering OR discovers OR discovery OR earliest OR early OR findings OR first OR foundation OR impacts OR institute OR latest OR new OR novel OR only OR original OR originally OR potential OR potentialities OR potentiality OR potentialize OR potentials OR predict OR rediscovery OR since)	查全率和查准率综合效果较好
摘要和引文拓展特征词	摘要 & 引文	@ Abstract(base OR based OR bases OR basis OR change OR changed OR changes OR changing OR conclusive OR confirming OR credible OR demonstrate OR essential OR evidence OR evidenced OR findings OR foundation OR grounds OR impacts OR important OR key OR major OR new OR novel OR potential OR potentials OR predict OR primary OR proof OR significant) OR @ Citation(affect OR beginning OR change OR changed OR changes OR changing OR contributions OR devote OR discover OR discovered OR discoverer OR discoveries OR discovering OR discovers OR discovery OR earliest OR early OR findings OR first OR foundation OR impacts OR institute OR latest OR new OR novel OR only OR original OR originally OR potential OR potentialities OR potentiality OR potentialize OR potentials OR predict OR rediscovery OR since)	查全率最高

参考文献：

[1] GARFIELD E. The 1976 articles most cited in 1976 and 1977. 1. Life sciences. [J]. Essays of an information scientist, 1979, 13(4): 81 -99.

[2] PONOMAREV I V, WILLIAMS D E, HACKETT C J, et al. Predicting highly cited papers: a method for early detection of candidate breakthroughs [J]. Technological forecasting and social change, 2014, 81(1):49 -55.

[3] HUANG Y H, HSU C N, LERMAN K. Identifying transformative scientific research[C]// 2013 IEEE international conference on data mining (ICDM). Melbourne: IEEE, 2013:291 -300.

[4] WOLCOTT H N, FOUCH M J, HSU E R, et al. Modeling time-dependent and -independent indicators to facilitate identification of breakthrough research papers[J]. Scientometrics, 2016,107(2): 807 -817.

[5] 杜建，孙轶楠，张阳，等. 变革性研究的科学计量学特征与早期识别方法[J]. 中国科学基金，2019, 33(1):90 -100.

[6] RADEV D R, AMJA D. Rediscovering ACL discoveries through the Lens of ACL anthology network citing sentences[C]// Proceedings of ACL 2012 special session on the 50th anniversary of ACL. Stroudsburg: Association for Computational Linguistics, 2012:1 -12.

[7] SMALL H, TSENG H, PATEK M. Discovering discoveries: identifying biomedical discoveries using citation contexts[J]. Journal of informetrics, 2017, 11(1):46 -62.

[8] SIOLAS G. Support vector machines based on a semantic kernel for text categorization[C]// Proceedings of the international joint conference on neural networks. Como: IEEE Computer Society, 2000: 205 -209.

[9] 刘丽珍，宋瀚涛. 文本分类中的特征选取[J]. 计算机工程, 2004, 30(4): 14 -15.

[10] SALTON G, BUCKLEY C. Term-weighting approaches in automatic text retrieval[J]. Information processing & management, 1987, 24(5): 513 -523.

[11] 谷俊，严明. 基于中文专利的新技术术语识别究[J]，情报科学，2013, 31(2):144 -149.

[12] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [EB/OL]. [2020 -02 -23]. https://arxiv. xilesou. top/pdf/1301.3781. pdf.

[13] CHEN C, SONG M, HEO G E. A scalable and adaptive method for finding semantically equivalent cue words of uncertainty[J]. Journal of informetrics, 2018, 12(1):158 -180.

[14] PONOMAREV I V, WILLIAMS D E, HACKETT C J, et al. Predicting highly cited papers: a method for early detection of candidate breakthroughs [J]. Technological forecasting and social change, 2014, 81:49 -55.

[15] Science newsletters[EB/OL]. [2020 -02 -23]. http://www.sciencemagchina. cn/highlights141219. aspx.

[16] Breakthrough of the year[EB/OL]. [2020 -02 -23]. http://en. wikipedia. org/wiki/Breakthrough_ of_the_Year.

[17] Colil[EB/OL]. [2020 -02 -23]. http://colil. dbcls. jp/browse/papers/.

[18] FUJIWARA T, YAMAMOTO Y. Colil: a database and search service for citation contexts in the life sciences domain[J]. Journal of biomedical semantics, 2015, 6(1):38.

[19] DING Y, ROUSSEAU R, WOLFRAM D. Text mining with the Stanford CoreNLP[J]. Replicable science of science studies, 2014(10):215 -234.

[20] 刘欣,余贤栋,唐永旺,等. 基于特征词向量的短文本聚类算法[J]. 数据采集与处理, 2017,32(5):1052-1060.
[21] PYYSALO S, GINTER F, MOEN H, et al. Distributional semantics resources for biomedical text processing[J]. Proceedings of languages in biology and medicine, 2013.
[22] KANERVA P, KRISTOFERSON J, HOLST A. Random indexing of text samples for latent semantic analysis[J]. Proceedings of the annual meeting of the Cognitive Science Society, 2000, 22(22): 1036-1036.
[23] CLEVERDON C. The cranfield tests on index language devices [J]. Aslib proceedings, 1967, 19(6):173-194.

作者贡献说明:
杨雪梅:突破性研究特征词识别方法优化及实现、论文撰写;
王雪:数据获取及预处理;
杜建:研究思路设计;
唐小利:优化研究设计及论文审核。

Identifying Feature Words Based on Abstracts and Citation Text Corpus of Breakthrough Research

Yang Xuemei[1] Wang Xue[1] Du Jian[2] Tang Xiaoli[1]
[1] Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100005
[2] National Institute of Health Data Science, Peking University, Beijing 100191

Abstract: [**Purpose/significance**] Based on the author's descriptive evaluation of his research and the critical citations of later researchers, the abstract and citation corpus of the breakthrough research are used to extract the feature words. Feature words can be used to understand the abstract and citation corpus features of the breakthrough research and contribute to the identification of breakthrough research. [**Method/process**] Key documents selected by Science as "Breakthrough of the Year" and "key publications" of Nobel Prize winners were selected as breakthrough research corpus data. Feature words were extracted by integrating abstracts and citation corpus of the paper. In the feature word extraction, the Stanford CoreNlp tool was used to perform word frequency statistics on the corpus, and the feature words were filtered in combination with expert opinions. Then we used the semantic relationship of medical texts to semantically expand feature words, which were used as the seed words. Finally, the retrieval and recognition effects of the abstract and citation feature words were further compared by the recall rate and the precision rate. [**Result/conclusion**] In the breakthrough research corpus, we selected 8 feature tokens of abstract corpora and 8 feature tokens of citation corpora. In the retrieval and recognition of feature words, the recall rate of the extended feature words of abstracts and citations is the highest, the precision of citation feature words is the highest. The comprehensive effect of the recall rate and precision of citation expansion feature words are better.

Keywords: breakthrough research feature words abstract text citing sentence